Clickbait intensity prediction and its relevance to Fake News

Information Retrieval and Extraction Project

Team 22 **Project 12** **Mentor: Vijaya Saradhi**

Palash Sharma (2019101082)	Pulkit Gupta (2019101078)	Aryan Jain (2019101056)	Archit Jain (2019101053)	Vemuri Murthy (2021900015)

Introduction

We need to predict the clickbait intensity of any news title/article and its relevance towards fake news. We shall make a suitable and reliable pipeline to report the clickbait intensity of a link and simultaneously predict to our best whether the article is real or fake.

Importance

Consuming news from social media is becoming increasingly popular nowadays, basically anyone, anywhere, can produce and help circulate content for other people to read. But this content on the other side also turned out to be fake, as there is no source verification, fact checking or accountability. Also, these fake news and clickbaits can quite easily affect the user's views and affect their decision making. So, it becomes conditionally important to identify clickbaits and fake news in the information shared with us.

Relevant documents

- In this paper, a system is presented to detect the stance of headlines with regard to their corresponding article bodies.This approach can be applied in fake news, especially clickbait detection scenarios.
- In this paper, multiple regression models have been used to assign a score to the links corresponding to their clickbait attributes with an 85% accuracy. This is a new approach against the traditional binary classification of links as clickbaits.
- This paper examines potential methods for the automatic detection of clickbait as a form of deception. It has methods for recognizing both textual and non-textual clickbaiting cues which are generated through surveys, leading to the suggestion that a hybrid approach may yield best results.
- This paper explains how to evaluate a model on the basis of different parameters like their accuracy, presion, recall and F1 score.
- In this paper, they propose various methods for clickbait intensity prediction based on the title of the post by using a benchmark dataset for evaluating multiple models.

- This paper comprises comparative analysis over explicit and implicit profile features performed between these user groups, which reveals their potential to differentiate fake news.

Plan

We are planning to predict the clickbait intensity of news articles from a fractional scale and whether the news articles are fake or not. We will then combine the results of these two to find the relevance between clickbait intensity and fake news.

We plan to try regression models such as Simple linear regression(LR), Ridge Regression (RR), Gradient Boosted Regression(GBR), Random Forest Regression(RFR), Adaboost Regression{ABR) for calculating clickbaiting(although some other classification methods have also been used before). Traditional methods perform a binary classification of a text being clickbait or not. More recent papers have also tried continuous distribution of clickbaiting between [0, 1] and our particular use case(linking clickbaits to fake news) will benefit with such a continuous range hence our implementation shall incline with the latter. We will check the results of the above mentioned algorithms and find the best suited one.

Some papers also mention ways to check the relation between a title and its contents and label each article as “related”(“agree,” “disagree”, “discuss,”) or “unrelated” by finding number of n-grams matching in the headline and the article and multiplying it with tf-idf.

For the second task of fake news detection, we intend to try the PA(Passive Aggressive) algorithm and KNN algorithm for classifying article content to be real or fake.

Finally, we shall combine the results in order to predict fake news related to clickbaits. We are not completely sure on what might work hence we might try different algorithms from the ones we have mentioned above and for combining both the results we will analyze the results of both the results and derive some relation between clickbait intensity and fakeness of a news article.

Dataset

We have currently identified a few datasets that we can use for this task

- Webis Clickbait Corpus 2017
- NELAGT2019
- NELAGT2020
- FakeNewsNet

** We will endeavour to incorporate more datasets into our approach as we move forward if necessary.*

Evaluation

We will calculate MSE(mean squared error) for clickbait intensity prediction and F1 score for fake news prediction.Using these scores we will predict the fakeness of news and the clickbait intensity.